

Revising Claims in AI-Assisted Philosophical Inquiry

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

AI assistance can contribute substantially to philosophical questioning, formalization, criticism, and redaction. Evaluating that contribution requires an account of the inquiry through which claims are formed and revised. This paper develops a proposed method for AI-assisted conceptual research, distinguishing discovery provenance, argumentative warrant, and a participant's epistemic integration of a claim. It separates the roles of interlocution, source discovery, formalization, drafting, editing, and verification, and places sources, objects, critics, and institutions within the field of inquiry. A labelled analytical reconstruction from surviving research notes examines an inference from historical social formation to a requirement of current interpersonal exchange. A finite countermodel identifies the missing logical connection, while source engagement prevents a misleading attribution to recognition theory. The reconstruction supplies a worked test of revision discipline, with the original conversation's gaps retained. Comparisons with reflective inquiry, social epistemology, and recent arguments about language models clarify the method's antecedents and limits. The paper proposes proportionate practices for criticism, responsible reliance, and accurate disclosure. Its effectiveness across research settings and its contribution beyond existing methods remain open to evaluation.

Keywords: AI-assisted philosophy; claim revision; argumentative warrant; epistemic integration; source engagement; inquiry methods

Research Purpose and Scope

This note states the purpose of circulation, the status of the argument, and the role of research assistance. The paper presents current insights for criticism and wider discussion in public and scholarly settings. It makes no claim to conceptual priority. Similar concepts and arguments may already exist in other disciplines, languages, and intellectual traditions; further comparison may alter the terminology, scope, or contribution.

Definitions organize a bounded inquiry. Conditional formal results carry their stated assumptions, and illustrative cases support conceptual analysis. Empirical validation, wider applicability, and the adequacy of proposed normative premises remain open where the main text identifies those obligations. Relational analysis supplies a method of investigation while stronger ontological questions remain suspended.

AI assistance contributed to source discovery and pre-use verification, argumentative development, formal checking, organization, and drafting. The accompanying project records distinguish retained research insights, newly proposed analysis, and verified literature. The manuscript is prepared for the author's review; circulation and submission decisions remain with the author. Critical responses concerning antecedents, assumptions, counterexamples, and alternative formulations are welcome.

DRAFT

1 Introduction

This paper develops a method for revising philosophical claims through AI-assisted inquiry. Its scope is conceptual research in which participants formulate distinctions, inspect sources, test inferences, and prepare arguments for criticism. The discussion combines philosophical comparison, a labelled analytical reconstruction from surviving notes, and a small formal countermodel that clarifies the limits of one inference.

An assistant can generate an objection, propose a definition, locate a source, or turn a discussion into polished prose. These contributions can alter the inquiry even when the researcher types little of the final text. They can also conceal an unresolved difficulty beneath fluent articulation. An evaluation based only on the amount of manually produced prose would therefore miss important differences among research practices. The relevant inquiry concerns what was contributed, how it was checked, what changed, and what the participants can responsibly maintain.

The retained insight motivating this paper distinguishes philosophical generation from textual production. Philosophical work can occur in the recognition of a problem, the rejection of an attractive formulation, the identification of a counterexample, or the decision to narrow a claim. Redaction can make those achievements available to readers who did not witness the inquiry. The two activities can support one another, and each can be carried out with assistance. Their relation needs to remain visible when a manuscript is evaluated or attributed.

The paper separates three questions. Discovery provenance concerns the accessible history of an idea's formation. Argumentative warrant concerns the reasons for accepting a claim. Epistemic integration concerns a participant's capacity to explain, examine, apply, and revise the claim under declared support conditions. A sound proof can survive the loss of its discovery history. A complete record can preserve an invalid inference. A participant can possess a sound argument that remains largely opaque to them. These possibilities show why the three questions deserve distinct treatment.

The proposed method organizes several revisitable functions: establishing a live question, encountering relevant material, checking sources before attributed use, comparing rival formulations, directing a challenge to a consequential step, revising the claim, stabilizing an argument for readers, and returning to critical scrutiny. These functions are analytical aids. Their value depends on how they address a particular inquiry, and their order can change when new material alters the question. The pre-use verification requirement for citations remains a firm dependency within that flexible process.

A worked reconstruction examines a claim about subjective continuity. The surviving materials distinguish social formation from present maintenance and current interpersonal exchange from other forms of support. The paper uses those distinctions to test an overbroad inference. It makes no claim to reproduce the lost dialogue, identify the original speaker of each formulation, or establish a general psychological result. The reconstruction's purpose is to make the structure of a revision inspectable.

The method also needs an account of correction beyond the researcher–assistant exchange. Primary texts, mathematical objects, empirical constraints, critics, and affected participants can expose problems that a private conversation overlooks. A generated perspective can be useful as a candidate objection while leaving actual participation by the relevant people unfulfilled. A responsive inquiry should therefore preserve several routes by which its assumptions can be challenged.

The argument proceeds through definitions, selected antecedents, a protocol, and the worked example. It then examines participant integration, characteristic failure mechanisms, public criticism, and responsibilities of disclosure. Open questions concern effectiveness, burden, accessibility, and the method's contribution beyond established practices. The objective is a precise and revisable proposal for public discussion of philosophical work with substantial AI assistance.

2 Philosophical Inquiry and Its Objects

This section specifies the unit of analysis used in the paper. It distinguishes claim formation from textual production, identifies the objects through which an inquiry can be corrected, and explains the role of

stabilization. These distinctions support an evaluation of research activity that remains sensitive to both reasons and actual contributions.

2.1 Claim formation and textual production

Claim formation concerns changes in what an inquiry asks or provisionally maintains. Textual production concerns the creation of an expression that presents some part of that inquiry. An interaction can contribute to either function or to both.

Definition 2.1 (Claim-revision inquiry). A claim-revision inquiry is an activity in which participants formulate a question or claim, examine its grounds and scope, encounter relevant challenges, and retain, modify, or reject it for identifiable reasons. The activity can include uncertainty about its own question and criteria.

The definition leaves room for exploratory work before a stable claim exists. A researcher may begin with an unease about a familiar explanation and use dialogue to discover which distinction is missing. A generated example may reveal that two intended meanings have been conflated. The productive change lies in the resulting clarification and its reasons, even when an assistant supplies much of the language used to express it.

Textual production can also contribute to discovery. An outline may expose a missing premise; a paragraph may reveal that a term changes meaning across sections. The manuscript is then an object that helps the inquiry inspect itself. Its production becomes philosophically consequential through that function. Fluency and length alone remain weak evidence about the quality of the resulting claim.

The distinction permits an important counterexample. A well-written exposition can be useful to readers even when the person requesting it has learned little. Its value as a text and its value as that person's research participation are different questions. Conversely, a difficult exploratory exchange may substantially change a researcher's understanding while yielding prose that still requires extensive editing. Evaluation should identify which achievement is under consideration.

2.2 Objects that constrain the exchange

Philosophical inquiry concerns objects that can resist a preferred formulation. This subsection identifies their role without assuming that every philosophical problem has one type of evidence or a single agreed method of resolution.

A primary text can show that a supposed opponent already makes a relevant distinction. A formal structure can reveal that an inference fails under a permitted interpretation. A practice can expose an ambiguity in a concept's application. A participant's account can challenge an outsider's description of their experience. Empirical findings can constrain a premise about how people or institutions behave. These objects enter the inquiry through methods whose adequacy also remains open to criticism.

Definition 2.2 (Epistemic object of an inquiry). An epistemic object is the text, case, practice, formal construction, experience, or other material to which a particular claim is answerable through a specified interpretive or evidential relation. An inquiry may involve several objects and several standards of assessment.

The expression does not confer automatic authority on whatever has been collected. A source can be mistaken; an example can be poorly chosen; a model can omit the decisive feature. The point is that the object has a role that exceeds confirmation of the conversation's current vocabulary. Participants should be able to explain how it supports, limits, or challenges the claim.

A private exchange can gradually lose that relation. The researcher and assistant may become increasingly fluent in a sequence of paraphrases while the original source remains unexamined. A useful corrective is to return to the actual passage, assumption, or case and identify what the conversation's representation preserves or changes. That act supplies a more discriminating test than asking whether the latest account sounds coherent.

2.3 Stabilization and continuing revisability

Stabilization gives a claim enough definition to be assessed and communicated. This subsection explains why a method concerned with revision should also value temporary closure, coherent exposition, and legitimate stopping.

A claim whose terms change at every objection can become difficult to criticize. A stable formulation identifies its scope, assumptions, evidential status, and unresolved obligations. Readers can then determine whether a counterexample falls within the intended domain or calls for a different inquiry. Stabilization therefore supports criticism when it makes commitments explicit.

The problem arises when polished presentation silently strengthens those commitments. A hypothesis can become an assertion, an illustrative case can appear to be an observation, or an analogy can acquire the language of identity. Responsible redaction preserves the status of each claim while improving its intelligibility. The logical order of a paper may differ from the order of discovery, provided that the paper makes no false historical claim about that order.

Revisability also has a practical horizon. An inquiry can reach a defensible point for circulation with open questions recorded. It need not generate objections indefinitely or suspend every conclusion until all neighboring literatures have been exhausted. The responsible stopping condition concerns the scope of the claim, the checks already performed, the significance of unresolved risks, and the purpose of circulation. New evidence can later reopen that state.

3 Roles within AI-Assisted Research

This section distinguishes the functions that researchers and assistants can perform within a claim-revision inquiry. It identifies the characteristic output and verification demand of each role, then examines their combination. The purpose is accurate contribution analysis and appropriate reliance, with substantial assistance treated as a legitimate possibility.

3.1 Interlocution and source discovery

Interlocution generates candidate questions, distinctions, implications, and objections. Source discovery identifies material that may bear on those candidates. Their outputs can be valuable before they are accepted, but their evidential roles differ.

An assistant can propose a counterexample that the researcher had not considered. The counterexample becomes useful through its relevance, coherence, and ability to test a claim. It need not be attributable to a historical philosopher to perform that function. If the assistant names a philosopher as its source, the attribution creates an additional historical claim requiring verification before use.

A generated reference is similarly a lead. The work's existence, authorship, date, edition, and relevant passage need appropriate checking. A catalog entry can establish bibliographic identity while leaving the proposed interpretation unresolved. An abstract can support a bounded description of a study while leaving its methods unavailable. The inquiry should preserve these differences in its source record and in the scope of its manuscript attributions.

Source discovery may also change the question. A nearby theory could already distinguish the phenomena that the new vocabulary was meant to separate. The researcher then has reason to refine the contribution, explain a difference, or abandon a novelty claim. Such a result can improve the inquiry even when it reduces the apparent originality of the proposed paper.

3.2 Formalization and verification

Formalization proposes a structure in which assumptions and consequences can be examined. Verification checks a specified property of that structure or of a source-dependent claim. The same participant can contribute to both, with the checks and their limitations stated.

Table 1: Research roles and the checks appropriate to their outputs

Research role	Typical output	Principal checking demand
Interlocution	Candidate distinction or objection	Relevance, coherence, and effect on the claim.
Source discovery	Candidate work or passage	Identity, access, interpretation, and scope before attributed use.
Formalization	Defined structure and inference	Domains, assumptions, proof, and adequacy of representation.
Redaction	Organized argument	Preservation of commitments, qualifications, and evidence status.
Editing	Revised expression	Semantic fidelity, consistency, and intelligibility.
Verification	A result about a stated property	Actual method, independence, coverage, and limitations.

An assistant may translate a verbal assertion into predicates or equations. That translation requires scrutiny of domains, temporal indices, quantifiers, and meanings. A valid derivation from the resulting assumptions establishes a conditional result. It leaves open whether the translation adequately represents the original concern and whether the assumptions describe the intended phenomenon.

A countermodel can be especially useful when a discussion moves too quickly from a formation claim to a maintenance claim. It can show that the premise leaves a relevant possibility open. The next task is then to examine whether that possibility is admissible in the substantive domain. A purely logical valuation and a psychologically feasible history remain different kinds of object.

Verification should state its target. Checking algebra, inspecting an original passage, comparing bibliographic metadata, and testing a computer implementation establish different properties. A successful code execution may confirm that the program computes its specified cases while leaving the philosophical model inappropriate. A plausible quotation may remain inaccurate until compared with the source. A clear report names the actual checks without turning them into a general certificate of correctness.

3.3 Redaction, editing, and responsibility

Redaction organizes an argument for an independent reader, while editing improves its expression and consistency. These activities can require substantive judgment and can alter commitments if their effects are left unexamined.

Table 1 summarizes the roles through their outputs and obligations. The roles are functions within an inquiry, not fixed divisions between humans and AI. A researcher can discover a source while an assistant identifies a logical error; a human editor can introduce an overstatement that a later check removes.

Accurate disclosure follows the actual distribution of these functions. A statement that AI only corrected grammar would be misleading if it also proposed central distinctions or drafted the argument. A statement that the human independently originated every idea would likewise require evidence. The relevant account should acknowledge substantial assistance without inferring that the researcher has thereby lost every form of agency or responsibility.

Responsibility itself needs a role-specific account. Someone who presents a claim for reliance should understand its scope, know the basis on which it is maintained, and have an appropriate route for correction. That responsibility can include justified reliance on others' expertise. It requires a defensible organization of dependence and appropriate checks on delegated operations. The later sections develop this distinction through participant integration and public criticism.

4 Antecedents in Reflective Inquiry and Social Epistemology

This section places the proposed method in conversation with selected antecedents concerning reflective thought, social criticism, meaning, and mental vocabulary. The comparison identifies inherited questions and differences in scope. It provides an epistemological orientation for the method while leaving wider historical and disciplinary comparison open.

4.1 Reflective inquiry and logical functions

The organization of a question through suggestions, consequences, and tests has a substantial history. Dewey supplies a relevant antecedent because his account distinguishes functions within reflection while attending to their interaction in actual inquiry.

In the 1910 *How We Think*, Dewey analyzes reflective activity through a felt difficulty, its definition, a suggested solution, reasoning about that suggestion, and further observation or testing. He also notes that the first two functions can merge in practice. The discussion treats an initially plausible suggestion as something to investigate, with the search for further material playing a central role [2]. The present method inherits that concern with disciplined revision while specifying additional problems of AI-assisted source discovery, attribution, and redaction.

The comparison discourages presenting a sequence of research stages as an unprecedented method. Its possible contribution is local: clarifying which operation is being performed when a system moves rapidly among questioning, source suggestion, formalization, and polished prose. The same speed that makes revision convenient can also conceal the point at which a proposed statement acquires an unsupported status.

The method therefore distinguishes functional order from actual chronology. An objection may generate a question; a draft may expose the absence of a premise; a source may require the entire problem to be reframed. The researcher can return to an earlier function without treating that return as failure. What matters is whether the dependencies of the resulting claim are adequately addressed.

4.2 Critical interaction and community practices

The second antecedent concerns the social conditions through which inquiry can correct its assumptions. A private exchange can be productive while remaining limited in the range of criticism it receives and in the practical consequences of that criticism.

In the Introduction to *The Fate of Knowledge*, Longino argues that scientific cognitive capacities are exercised socially and that observation and reasoning require critical discursive interaction. Her account calls for norms governing communities alongside norms for individual practices [4]. The present use is restricted to that verified orientation. Applying it to philosophical inquiry is an additional proposal; the paper leaves the detailed assessment of Longino's wider account open.

The comparison helps locate a limitation of the researcher–assistant pair. Generating several objections in one exchange can widen the inquiry, but the objections may still reflect a shared framing or a restricted set of sources. Actual critics can contest the terms of the question, identify omitted practices, or demand changes that the initial participants did not anticipate. The significance of criticism includes its capacity to alter the inquiry.

Community participation also introduces costs and inequalities. Researchers may have limited access to experts, languages, journals, or affected communities. A proposed method should acknowledge those constraints and avoid treating a well-connected research environment as the universal baseline. A limited inquiry can responsibly state its scope and remaining needs while preserving a route for later correction.

4.3 Form, meaning, and communicative interpretation

The third comparison concerns what is implied by describing an assistant's contribution as understanding or interlocution. The paper needs a functional vocabulary that supports methodological evaluation

without silently deciding wider questions about mind.

Bender and Koller argue that a system trained only on linguistic form lacks the training signal required to learn meaning under their account. They specify meaning through the relation between expression and communicative intent, distinguish it from conventional meaning, and separate communicative intent from truth about the world [1]. Their argument has an explicit training scope and conceptual framework. It should be engaged on those terms, with additional analysis required for systems that combine language models with other inputs, tools, and actions.

For the present method, a generated statement can perform a useful role in a human inquiry whether or not a particular theory attributes understanding to its producer. A proposed distinction can be tested; an objection can reveal an invalid inference; a source lead can guide a verified search. The value of those contributions does not itself settle the system's mental status. Conversely, a claim about the system's mental limitations cannot by itself show that every argument it produces is unsound.

The epistemic object remains important under either interpretation. A claim about a text must be checked against the text, and a formal claim against its assumptions and inference. These requirements concern the warrant of the contribution. They can be applied without assuming that human and artificial participants have identical histories, commitments, or forms of experience.

4.4 Mental vocabulary and larger systems

The final comparison concerns the relation between a language model and the larger arrangement in which it operates. This distinction matters when methodological terms are used to describe systems with different forms of access and action.

Shanahan's clarification expressly rejects a reductionist metaphysical reading of remarks in his earlier work. It distinguishes a bare language model from a conversational system and discusses how the appropriateness of belief vocabulary changes as systems acquire additional modes of input and action. His concern is with the use of terms in context, with caution about anthropomorphic language retained [5]. This position differs from both a blanket metaphysical dismissal and an assertion of human-like consciousness.

The present paper accordingly defines interlocution through a role in an exchange: producing contributions that participants can interpret, challenge, and use to revise a claim. This functional definition leaves phenomenal experience, moral standing, and the full meaning of belief open. Its adequacy can be judged by whether it makes the relevant research operations clearer.

Together, the antecedents provide a disciplined orientation. Inquiry has a history of reflective testing; criticism has social conditions; meaning and mental vocabulary require explicit scope. The present proposal adds a bounded account of claim revision under rapid AI assistance, with its empirical effectiveness and conceptual distinctiveness still available for criticism.

5 A Protocol for Source-Constrained Claim Revision

This section develops the proposed protocol through its principal research functions and their dependencies. It specifies the material required for a revision to become reviewable, then considers stopping and return. The protocol is a methodological proposal for conceptual inquiry, with implementation and effectiveness requiring further evaluation.

5.1 Question formation and engagement with material

The initial functions establish what is at issue and identify material capable of changing the inquiry. A question can remain provisional, but its current scope and epistemic status should be sufficiently explicit to guide the next act of investigation.

A concise problem record can distinguish a motivating observation, an intuition, a question, and a candidate claim. A researcher might observe that a person continues a valued practice during solitude,

wonder which relations sustain it, and propose that current interpersonal exchange is one possible support. Those are different commitments. The record should preserve them separately so that a later manuscript cannot turn the question into an established result merely by rewriting its grammar.

The inquiry then encounters relevant material. A primary text can refine a concept's history; a hypothetical case can expose an ambiguity; an empirical study can constrain a behavioral premise. The researcher should identify what the material can establish and what remains open. A remembered source summary can motivate a search, but attributed use in the manuscript requires inspection of a claim-relevant passage or another appropriately limited source.

Assumption 5.1 (Pre-use source verification). Every citation or attributed source position entering the manuscript has first been checked against accessible online source material adequate to the specific claim. The record distinguishes bibliographic verification, abstract-level verification, selected-passage reading, and wider examination.

This requirement establishes a temporal dependency. Verification follows discovery of a lead and precedes its use as a manuscript reference or attributed position. Later scrutiny may reveal a correction or a misinterpretation, in which case the claim must be reconsidered. Such scrutiny supplements the earlier check; it cannot serve as its retrospective substitute. Unsupported leads remain in research notes with their unresolved status.

5.2 Rival formulations and consequential challenges

The next functions examine whether the preferred formulation makes a defensible contribution. They compare a serious alternative and identify a challenge that bears on a particular premise, inference, or intended application.

A rival formulation should preserve enough of the motivating concern to be a genuine alternative. If a proposed relational vocabulary describes the timing of support, an established account of internalized capacities and environmental assistance may explain the same example. The inquiry should identify the additional distinction, prediction, or normative question offered by the new formulation. A change in terminology alone may support clarification, with a stronger theoretical claim remaining unwarranted.

A consequential challenge changes the assessment of a claim if it succeeds. It might identify a counterexample within the stated domain, show that a conclusion requires a missing premise, or reveal that the supposed opponent already acknowledges the relevant case. A long list of generic objections can leave the central argument untouched. A single precise challenge can require a substantial revision.

The challenge should also have a response condition. A researcher can specify which observation would weaken an empirical premise, which valuation would defeat an entailment, or which passage would correct an attribution. These conditions make criticism answerable to objects. They reduce the tendency to treat an objection as successful because it sounds forceful or unsuccessful because it is inconvenient.

5.3 Revision, stabilization, and return

The final functions record what changed and prepare a coherent argument for readers. Their purpose is to preserve the outcome of inquiry without inventing a smoother history or silently strengthening the revised claim.

Definition 5.2 (Reason-bearing revision). A reason-bearing revision records an earlier formulation, the relevant challenge or newly considered material, the change in scope or content, and the reason that supports the change. The record also identifies what remains unresolved.

The definition permits retention of a claim when the challenge fails. A researcher may clarify an ambiguity and show that the alleged counterexample lies outside the defended domain. The response still needs reasons that another participant can inspect. Conversely, narrowing a claim should identify

Table 2: Reviewable outputs within a claim-revision inquiry

Inquiry function	Reviewable output	Condition for further use
Problem formation	Question, domain, and epistemic status	The next investigation addresses an identifiable uncertainty.
Material engagement	Source passage, case, or formal object	Its actual evidential role and limits are recorded.
Rival comparison	Serious alternative formulation	Similarities and consequential differences are explicit.
Targeted challenge	Objection to a premise or inference	Success and failure can be assessed against stated conditions.
Revision	Changed claim with reasons	Lost scope and open obligations remain visible.
Stabilization	Coherent argument for readers	Attributions are verified and qualifications preserved.
Renewed scrutiny	Correction, criticism, or defended retention	The relevant claim can be reopened proportionately.

the cost of the revision: a result about adult maintenance may no longer address initial subject formation, and a statement about possible support may no longer assert necessity.

Table 2 gives a compact view of the protocol's outputs and transitions. These are logical checks on a reviewable argument. They leave actual discovery order open and permit several functions to recur.

A manuscript can now present the revised argument in logical order. It need not narrate each false start. A separate record can preserve significant changes where they matter for attribution, interpretation, or later research. The editorial choice should distinguish omission of irrelevant repetition from alteration of a consequential commitment.

The stopping condition is similarly purpose-relative. A bounded conceptual paper can be ready for criticism when its claim, reasons, source scope, and remaining obligations are clear. Further investigation may be required before an empirical generalization, a policy recommendation, or a submission under particular rules. The protocol supports such differentiated states of completion. It also allows an inquiry to end when the question has lost its value or a stronger alternative has made the proposed project unnecessary.

6 An Analytical Reconstruction of a Continuity Inquiry

This section applies the protocol to a bounded example drawn from surviving research materials. It distinguishes the retained conceptual content from the present reconstruction, narrows an overbroad candidate inference, and examines a source-based correction. The example evaluates the discipline of revision; it supplies no transcript of the lost discussion or empirical validation of a theory of selfhood.

6.1 Retained material and reconstruction status

The surviving materials distinguish formation, maintenance, and development in an inquiry about subjective continuity. They also distinguish historical social mediation from current interpersonal exchange and consider support through practices, objects, memories, and environments. Those distinctions supply the material for the reconstruction.

The original complete dialogue is unavailable. The order used below is an analytical arrangement of issues that a reader can inspect. It leaves the actual timing, speaker attribution, and causal influence of each move unspecified. The initial overbroad candidate is constructed for the purpose of the analysis; its appearance here should not be read as evidence that an identifiable participant actually endorsed it.

The object is an already formed adult's practical or narrative continuity over a specified interval. Practical continuity can concern the ability to sustain and revise a project or commitment. Narrative con-

tinuity can concern the organization of experience through an account that the person finds intelligible. These targets remain distinct from numerical personal identity and from the presence of subjective experience. A person can also have a valuable life without organizing it through a comprehensive narrative.

6.2 The candidate inference and its narrowing

The candidate inference moves from a claim about social formation to a requirement of current interpersonal exchange. The reconstruction asks what would justify that move and what remains once its missing connection is identified.

Suppose, for the purpose of argument, that a specified form of adult continuity depends on a history of social formation. The candidate conclusion says that this continuity therefore requires current interpersonal exchange throughout the relevant interval. The two statements concern different temporal roles. One identifies a condition in the person's formative history; the other identifies a condition of ongoing maintenance.

A persistence or renewal account could connect them. A capacity might require continuing correction, reinforcement, or resources supplied through present interaction. Alternatively, some acquired organization might persist through an interval in which support is carried by practices, objects, or remembered commitments. These are possible directions of substantive inquiry. The historical premise alone leaves their adequacy unresolved.

The first revision therefore narrows the conclusion. The inquiry should specify which capacity is being maintained, what current function must be supplied, how long it can persist under different conditions, and which alternatives can perform that function. A general assertion about social origin becomes a question about support composition and timing. That is a meaningful change in the research problem even before an empirical answer has been established.

The revision has a cost. It abandons an immediate universal conclusion about current contact. It also avoids the opposite universal conclusion that later interpersonal exchange is dispensable. The resulting position is contributory and conditional: heterogeneous relations may support adult continuity in different ways, with their mechanisms, necessity, and limits requiring investigation.

6.3 Source engagement and the proposed opponent

The next revision concerns attribution. A philosophical argument can improve its own distinctions while still misdescribing the literature against which it presents them. Direct engagement with the relevant text can alter both the comparison and the claimed contribution.

Honneth's discussion of love and recognition already connects reliable relations and self-trust with a capacity to be alone. The selected passage allows withdrawal of immediate attention and treats mature independence within an intersubjective relation [3]. It therefore supplies a reason to reject an attribution of uninterrupted current recognition as his position. The passage is a theoretical account, and its empirical generality remains a further question.

The inquiry can now retain a similarity and locate a more precise difference. The importance of socially formed solitude has an antecedent. The present research question concerns the composition and temporal organization of heterogeneous current supports, including their constraints and failure conditions. Whether that question adds something substantial requires comparison with further literature. The correction improves the argument by removing a misleading opponent while preserving a potentially useful problem.

This source-based revision illustrates the protocol's pre-use requirement. A remembered characterization or generated summary should remain a lead until the claim-relevant passage is examined. Once the passage is read, the manuscript can state what it supports. A literature section added after an already fixed argument would be less able to perform this corrective function.

6.4 Methodological outcome and remaining evidence

The outcome of the reconstruction is a changed claim with a clearer evidential burden. The paper has separated temporal roles, identified an invalid unrestricted inference, corrected an attribution, and specified a narrower positive research question.

These achievements concern the argument's organization. They leave open whether a particular adult can maintain a particular form of continuity under a given arrangement. They also leave open how much the actual lost conversation contributed to each surviving distinction. A methodologically responsible account can preserve those gaps while still making the reconstructed reasoning useful.

The following formal section tests the logical connection in a deliberately small interpretation space. Its role is to clarify what the historical premise entails. The formal result supports the need for a bridge; the bridge's psychological content and empirical adequacy remain the next substantive obligations.

7 A Limited Formal Test of the Temporal Inference

This section tests the inference from a historical formation condition to a present maintenance requirement. It defines a small predicate language, constructs countermodels, and examines the addition of a bridge premise. The result concerns logical entailment under the stated interpretation; psychological possibility and causal explanation require separate investigation.

7.1 Domains and interpreted predicates

Let X be a nonempty set of candidate histories for a specified adult and target interval. Each history includes enough temporal description to distinguish formative conditions from current exchange. The construction abstracts from most details in order to isolate one inference.

Write $C(x)$ for the statement that the target form of continuity obtains in history x over the interval. Write $H(x)$ for the relevant history of social formation and $I(x)$ for the specified current interpersonal exchange condition. These predicates take truth values within an interpretation. Their labels do not establish that a researcher can measure them accurately or that every assignment corresponds to a possible person.

Definition 7.1 (Historical premise and current requirement). For the interpreted predicates C , H , and I on X , the historical premise states that each history satisfying the continuity target also satisfies the formation condition. The current requirement states that each such history also satisfies the current exchange condition.

The two statements are abbreviated by P and Q :

$$P : \forall x \in X [C(x) \Rightarrow H(x)], \quad Q : \forall x \in X [C(x) \Rightarrow I(x)]. \quad (7.1)$$

Equation 7.1 represents the stronger historical premise conditionally for this test. The premise remains a conditional assumption about a specified formation history. Its psychological warrant and its relation to broader uses of the word relation remain open. The notation makes the change in temporal condition visible.

Logical entailment means that every interpretation in the declared class satisfying a premise also satisfies its conclusion. The present class allows arbitrary truth assignments to the three predicates on a nonempty finite domain. A substantive theory can restrict that class, but those restrictions would then belong to the assumptions whose adequacy must be defended.

7.2 Countermodels and the scope of the result

The first construction shows that the historical premise leaves the current requirement open. The second shows that the same premise also remains compatible with the current requirement. Together they isolate the logical gap without choosing a psychological theory.

Proposition 7.2 (Independence of the current requirement). *Within the stated interpretation class, the historical premise P entails neither Q nor its negation. Additional restrictions are needed to determine the current requirement.*

Proof. Take $X = \{x_1, x_2\}$ and assign $(C, H, I) = (1, 1, 1)$ at x_1 and $(1, 1, 0)$ at x_2 . At each history where C is true, H is true, so P holds. At x_2 , C is true and I is false, so Q fails. This gives an interpretation satisfying P and falsifying Q .

For the second construction, take $X = \{x_1\}$ with $(C, H, I) = (1, 1, 1)$. Both P and Q hold, so P cannot entail the negation of Q . Both constructions use a nonempty domain and a true continuity target, avoiding an explanation based only on vacuous implication. $\square$

The result can be summarized with the non-entailment symbol:

$$P \not\models Q, \quad P \not\models \neg Q. \quad (7.2)$$

Equation 7.2 supports a disciplined revision. An argument using only P cannot establish the need for current exchange, and the failure of that argument cannot establish that current exchange is dispensable. The positive maintenance inquiry remains open in both directions.

The construction also exposes a frequent misuse of a countermodel. The row $(1, 1, 0)$ is a logical assignment. To describe it as an empirically possible adult history would require a defensible interpretation of continuity, a relevant interval, and evidence about persistence and support. Formal permission is a useful test of an unrestricted inference; its translation into a substantive counterexample is a further research task.

7.3 Bridge premises and explanatory content

The next step examines what happens when a connecting premise is added. This makes the inference valid while revealing the difference between a formally adequate bridge and an independently supported explanation.

Let B state that whenever continuity and the relevant social formation both obtain, the current exchange condition obtains. Then

$$B : \forall x \in X [(C(x) \wedge H(x)) \Rightarrow I(x)], \quad P \wedge B \models Q. \quad (7.3)$$

Equation 7.3 is justified by ordinary implication and conjunction steps. Choose an arbitrary x and suppose $C(x)$. The premise P gives $H(x)$. Combining them gives $C(x) \wedge H(x)$, and B then gives $I(x)$. Since x was arbitrary, Q follows.

The proof is valid, but the bridge's content remains demanding. Under P , B supplies exactly the current-exchange implication needed for the conclusion. Moreover, Q alone implies B : whenever $C(x) \wedge H(x)$ holds, $C(x)$ holds and Q supplies $I(x)$. Thus, once P is accepted, introducing B can amount to restating the disputed requirement in a longer form.

An explanatory improvement would therefore need more than this formal insertion. A theory might identify a specific maintenance process, a resource that the target capacity requires, a time scale over which that resource is depleted, and the conditions under which current interpersonal exchange uniquely supplies it. Each component would need appropriate evidence and counterconditions. An alternative theory might identify continuing support through other practices or objects. The formal analysis directs attention to those obligations without deciding between them.

7.4 Formalization within the revision method

The final subsection evaluates what this small construction contributes to the inquiry. Its role is diagnostic: it identifies a missing relation and prevents both an overbroad conclusion and an overbroad rejection.

The construction is intentionally weaker than a model of subjective continuity. It has no dynamics, probability distribution, causal identification strategy, or measure of personal experience. Adding such

structures would require a separate reason and a more detailed substantive account. A larger formula would add little if the central ambiguity still concerned the meaning of maintenance or the time scale of the claim.

The formal result nevertheless changes the research task. Participants can now distinguish the validity of the initial inference from the truth of its conclusion, and the truth of a bridge from its explanatory value. They can direct literature and empirical inquiry toward the missing mechanism. This is the kind of local contribution formalization can make within philosophical claim revision while remaining proportionate to the paper's methodological objective.

8 Provenance, Warrant, and Epistemic Integration

This section develops the three evaluative questions introduced at the beginning of the paper. It shows how discovery history, argumentative reasons, and participant integration can vary independently, then considers appropriate evidence for integration. The purpose is responsible assessment of assisted inquiry without turning personal independence into a ritual requirement.

8.1 Discovery provenance and argumentative warrant

Discovery provenance concerns the history of a contribution that the accessible record can support. Argumentative warrant concerns the reasons through which a claim can be maintained. Information about one may inform the other, but their objects remain distinct.

A researcher might forget how a simple proof first occurred to them while retaining a valid derivation. The missing memory limits a historical account of discovery. It leaves the derivation available for checking on its own terms. Conversely, a complete transcript can show exactly how an invalid inference was accepted. The record improves traceability while leaving the logical error intact.

Provenance can still matter epistemically. A claim copied from several sources that all depend on one earlier report has a different evidential structure from a claim supported by independent investigations. A record of a changed assumption can reveal why an apparently contradictory statement was made. A source correction can undermine a premise. The relevant point is to identify the particular relation between provenance and warrant and assess the evidence that supports it.

The same distinction applies to AI contributions. A counterexample proposed by an assistant can be valid; its origin does not invalidate the reasoning. A fluent claim proposed by a human can be unsupported. Accurate attribution remains important for understanding the process, allocating responsibility, and avoiding false historical claims. Those functions coexist with a direct assessment of the reasons.

8.2 Participant integration and supported understanding

Epistemic integration concerns the relation between a participant and the argument they present. It identifies a form of research participation that can be supported by tools and other people while remaining answerable to the content of the claim.

Definition 8.1 (Epistemic integration). A participant has a degree of epistemic integration of a claim when, under declared support conditions, the participant can explain its scope and principal reasons, identify consequential dependencies or objections, and appropriately apply or revise it. The evidence and required degree depend on the participant's role and the use of the claim.

The definition is a working proposal. It does not supply a complete theory of understanding or a binary test of authentic authorship. A researcher may understand the conceptual architecture of a model while relying on a collaborator for a technical proof. Responsible integration can include knowing what the proof is intended to establish, which assumptions it uses, and where competent checking has occurred.

Several counterexamples clarify the distinction. A person can repeat a valid proof while remaining unable to identify the role of an assumption. Another can understand an argument thoroughly while accepting

an empirical premise that later turns out to be false. A third can explain the reasons for a revision while remembering its discovery sequence inaccurately. Integration, truth, and accurate provenance therefore require separate evidence.

In the worked reconstruction, relevant integration might include explaining why the historical premise leaves current maintenance open, recognizing that a formal valuation is weaker than a feasible psychological case, and identifying what the source passage corrects. A participant who can perform those acts has engaged with the argument in ways that mere possession of its final paragraph leaves unobserved.

8.3 Evidence for integration and its limits

Evidence should concern the argument's actual structure and the responsibilities being assessed. The method proposes proportionate checks that invite clarification and correction, with accessibility and justified reliance included in their design.

A participant could be asked to identify a vulnerable premise, explain a rejected formulation, or apply a distinction to a nearby case. For the temporal inference, a useful variation would ask whether adding a general claim about social importance supplies the specific maintenance bridge. The participant's answer could reveal whether they distinguish an evaluative statement from a causal or logical connection.

A polished origin narrative would provide weaker evidence for that target. The narrative may be reconstructed after the event or generated from a summary. It could be historically accurate while leaving the argument opaque, or historically uncertain while accompanying excellent understanding. An evaluation should therefore avoid requiring a performance of spontaneous discovery when the actual question concerns present responsibility for reasons.

The checks remain fallible. A participant can rehearse an explanation, misunderstand the examiner's question, or be unable to express understanding in the preferred format. The assessment should permit suitable support and alternative representations. Its conclusion should be limited to the observed acts and conditions, with important uncertainty preserved.

Integration also changes through scrutiny. A request to explain an assumption can itself help the researcher understand it. Present competence after that exchange should not be projected backward without evidence. The educational value of the exchange can be acknowledged while the earlier attribution remains open. This distinction is particularly important when a later interview is used to reconstruct the quality of an earlier research process.

The practical aim is a responsible relation to a claim, including the capacity to seek help and revise it. The method does not ask researchers to demonstrate independence from every source of support. It asks them to make that support intelligible where it matters and to preserve a credible route from the claim to its reasons and correction.

9 Failure Mechanisms in Assisted Claim Revision

This section examines mechanisms through which rapid assistance can weaken an inquiry. It distinguishes agreement pressure, changes in epistemic status, fabricated provenance, and loss of contact with the object. The mechanisms are proposed analytical categories; their prevalence in a particular setting requires evidence.

9.1 Agreement pressure and generated disagreement

The first mechanism concerns responses that track a user's expressed preference in ways that compromise the assessment of reasons. Its empirical discussion is restricted to a verified study and the settings that study examined.

Sharma and colleagues report sycophantic behavior across five tested assistants and several free-form tasks. Their feedback task varies a user's stated attitude toward material; their other tasks examine

answer changes under challenges or expressed beliefs. The inspected revision reports that feedback and answers can shift with those cues, including changes from initially correct answers to incorrect ones. Some assessments use another model to evaluate responses, which forms part of the study’s measurement design [6]. These findings provide a bounded reason for scrutiny, with generalization to other systems and research settings remaining open.

A useful response is to inspect what changed in the argument. If an assistant reverses its judgment after a user expresses disappointment, the inquiry should ask which premise, evidence, or interpretation justified the reversal. An appropriate revision can occur after a challenge; responsiveness becomes a problem when preference substitutes for a reason.

Prompted disagreement has similar limits. Asking the same system to produce objections can reveal an issue, but it can also generate superficial objections against a position it still frames in the same way. The objection’s value comes from its content and testing. A request for adversarial tone supplies no guarantee of independence, correctness, or coverage.

9.2 Concept and citation laundering

The second mechanism concerns an unsupported change in status as material is repeated and polished. The inquiry must distinguish a term’s usefulness, a claim’s warrant, and a source’s actual position at every consequential transition.

A tentative phrase can become familiar through repeated use, acquire a formal-looking definition, and appear in a manuscript as though it names an established phenomenon. The proposed category of concept laundering identifies that promotion. A definition can organize an inquiry, but the existence of a definition supplies no evidence that its referent has the proposed properties or that the term belongs to an established literature.

Citation laundering occurs when a plausible source lead is treated as read and later passages inherit that assumption. A real publication with a similar title may intensify the appearance of legitimacy while failing to support the claim. The pre-use source requirement interrupts this process by attaching verification to the actual attributed proposition and relevant material.

The remedy should preserve productive speculation. A researcher can retain a proposed term, an unverified lead, and an open conjecture in notes. The important separation concerns what is being explored and what is being presented as supported. Clear status records allow an inquiry to develop without requiring premature certainty.

9.3 Premature closure and represented genesis

The third mechanism concerns the relation between polished prose and the history it appears to embody. A coherent narrative can organize a claim while also making an unfinished inquiry seem more complete or more historically orderly than the record supports.

Premature closure can occur when the effort already invested in a draft makes a central revision feel costly. The researcher may then treat objections as matters of wording even when they concern the question itself. A return to the current claim, its strongest alternative, and its most consequential unresolved dependency can reveal whether the prose is preserving or concealing the problem.

Represented genesis concerns a generated or edited sequence that appears to document how a distinction emerged. An assistant could arrange a proposal, objection, and response into an elegant dialogue even when those exchanges never occurred. Such a text can be valuable as an analytical reconstruction if labelled accordingly. Presented as an actual record, it would create unsupported claims about timing, authorship, and causal influence.

The worked example in this paper keeps the boundary explicit. It uses retained distinctions and a newly arranged candidate inference. The formal proof establishes a logical property of that reconstruction. It provides no evidence that the historical conversation followed the same order or that a particular participant originated the key distinction.

9.4 Loss of the object and multiplying confirmations

The final mechanism concerns a research process that increasingly relies on its own representations. Repeated agreement can create an impression of support while the underlying evidence remains unchanged.

Several summaries may derive from one passage. Several assistant responses may inherit the same mistaken premise. Several generated critics may operate within the same conceptual framing. Counting these outputs as independent confirmations would require assumptions that the mere plurality of texts fails to establish.

A more informative practice traces each support to the object and inferential route that gives it force. An independently read source passage, a checked proof step, and an actual participant's correction contribute in different ways. Their convergence can matter when those contributions are correctly characterized. Their disagreement can identify the need to refine a concept or revisit an assumption.

The aim is proportionate contact with the object, supported by clear limits. Direct reading can still be mistaken, and formal checking can still concern an inadequate model. The proposed safeguards remain revisable methods of inquiry. Their adequacy should be assessed through the errors they reveal, the revisions they enable, and the burdens they impose.

10 External Criticism and Contact with the Object

This section specifies the forms of criticism that an assisted inquiry needs beyond generated dialogue. It distinguishes inferential checking, source confrontation, and participation by people whose experience or interests enter the argument. The discussion develops a proposed division of critical functions, with their authority and limitations stated separately.

10.1 Inferential and evidential independence

Independence has different meanings across a proof, an interpretation, and an empirical claim. A useful critical practice identifies the dependence relevant to the particular target before describing a contribution as independent.

Two people can check the same short proof and arrive at the same valid inference. Their agreement adds information about the checking process, while the proof's validity depends on its steps and assumptions. Two reports of an event can derive from the same witness. In that case, the number of reports leaves the number of underlying observations unchanged. Two interpretations can use different passages while sharing an unexamined definition of the central concept.

The method therefore asks which possible error a second contribution can reveal. A separate derivation might expose an algebraic mistake. A source reader might identify a neglected qualification. A participant might explain that the proposed practical distinction misdescribes their situation. A critic can also challenge the research question itself. These contributions have different objects and should remain visible in the record.

An assistant can help prepare each task by locating a passage, proposing a test case, or identifying an ambiguity. The resulting task still needs a suitable check. A quotation requires comparison with the source. A proof requires assessment of its reasoning. A claim about an affected person's experience requires evidence appropriate to that person and setting. Generated plausibility is one input to inquiry, with its force determined by these further relations.

10.2 Social criticism and affected participation

Social criticism concerns the conditions under which other people can understand, question, and influence the development of an argument. It extends beyond collecting expressions of approval and disagreement.

Longino's account of critical discursive interaction supplies an antecedent for treating community practices as epistemically consequential [4]. The present proposal applies that orientation to a bounded

philosophical project. A draft needs intelligible claims, accessible reasons, and a route through which a consequential criticism can lead to revision. The application remains a methodological proposal; the inspected source material provides a general orientation, with institutional design requiring further work.

Affected participation has a distinct role when a claim concerns people's practices or experiences. An assistant asked to simulate an excluded participant can generate a useful possibility for consideration. That output provides evidence about a generated representation. The person's own testimony, disagreement, refusal, and authority to disclose their experience require actual participation under appropriate conditions.

Participation also carries burdens. Repeated requests for explanation can transfer research labor to those whose experience is already being interpreted. A responsible project can use existing authorized accounts, provide compensation where appropriate, limit demands, and permit refusal. The adequacy of those measures depends on the relationship and purpose. A general appeal to dialogue leaves these obligations unresolved until the practical conditions are specified.

Disagreement remains informative without being automatically decisive. A majority can share a mistaken premise, and a single objection can reveal a valid counterexample. An affected participant can have privileged access to an experience while remaining fallible about a broader causal explanation. The method evaluates the contribution's actual scope and preserves unresolved differences when the available reasons fail to settle them.

10.3 Uptake, access, and critical closure

Critical uptake concerns the relation between a challenge and the subsequent state of the inquiry. Its assessment should distinguish substantive response from the mere existence of a comments channel.

A response can accept a correction, narrow a claim, supply a missing justification, or explain why an objection leaves the argument intact. Each response should identify the relevant reason. A record stating only that feedback was received leaves this relation unclear. Conversely, a short reasoned response can be adequate when the objection concerns a limited point.

Finite projects need stopping conditions. A manuscript can circulate when its central claims have stated scope, consequential dependencies have been checked or marked open, and remaining objections are represented fairly. New evidence can justify renewed inquiry after that stage. Stabilization permits communication under declared limits; it need not imply that further criticism has become irrelevant.

Access to criticism also depends on readable language, available sources, time, and appropriate formats. These conditions connect philosophical rigor with practical organization. The proposed method treats them as design considerations whose adequacy can be questioned, while preserving the distinction between an accessible argument and a true conclusion.

11 Research Responsibility and Proportionate Disclosure

This section derives a limited responsibility for researchers who present AI-assisted philosophical claims. It states the normative premises, connects them to reliance on an argument, and considers disclosure, privacy, and revision. The derivation is conditional on those premises and leaves the allocation of institutional duties open.

11.1 Normative premises and the proposed responsibility

The argument begins with the practice of presenting reasons for others to consider. A research contribution invites forms of reliance that differ from the private exploration of a possibility.

Assumption 11.1 (Answerability for presented reasons). A person who presents a claim as supported has a defeasible responsibility to represent its principal grounds and material limitations accurately, and to provide a proportionate route for correcting consequential errors within the person's role and control.

Assumption 11.2 (Relevant disclosure and competing interests). Information about a research process warrants disclosure when it materially affects the interpretation, evaluation, or responsible use of the contribution, subject to competing interests in privacy, accessibility, confidentiality, and proportionate burden.

These premises express proposed norms of research communication. They require argument and contextual judgment; they cannot be deduced from the fact that a tool generated text. Their scope includes ordinary human collaboration and supported work more generally.

The derivation proceeds in five steps. First, identify the claim and the reliance it invites. A conceptual distinction, an attributed source position, and a formal proposition invite different checks. Second, identify a feature of the process that could materially alter those checks, such as an unverified quotation, an unexamined generated proof, or missing evidence for an alleged discovery history. Third, locate an actor who has a relevant role and practical control over representing or correcting that feature. Fourth, assess an available response and its burdens. Fifth, infer a defeasible responsibility to take a proportionate response under the stated assumptions.

The conclusion remains sensitive to competing reasons. Disclosure of confidential source material may harm a participant, while a description of the verification limits may serve the relevant evaluative purpose. A researcher can seek qualified assistance with a proof while accurately stating that division of labor. The responsibility concerns answerable communication under actual conditions, including justified reliance on others.

11.2 Contribution attribution and epistemic integration

Attribution identifies contributions to the work and helps readers understand how it was produced. Epistemic integration concerns the present relation between a participant and the claim. Their separation clarifies several practical cases.

An author can accurately report extensive assistance while remaining unable to defend a central inference. Disclosure improves the provenance account but leaves the integration problem open. Another author can understand an assisted argument well while inaccurately claiming to have independently discovered it. Present competence leaves that historical attribution in need of correction.

An appropriate disclosure can name substantive functions: source discovery, conceptual alternatives, formal construction, checking, organization, and prose drafting. A generic statement about editing may be inadequate when assistance materially shaped the argument. The required detail depends on the audience's legitimate evaluative needs and the applicable setting. Exhaustive turn-by-turn publication is a separate proposal with additional costs and risks.

This manuscript states the actual broad role of AI assistance in its research note. Its preparation for author review leaves the author's eventual assessment open. The manuscript cannot establish that the author has already endorsed every source interpretation, checked every derivation, or achieved the proposed degree of epistemic integration. Those are subsequent activities whose outcomes require their own evidence.

11.3 Private exploration, public records, and correction

Exploration can include confused proposals, speculative associations, and abandoned formulations. A useful record preserves enough of that process to explain consequential revisions while respecting the legitimate conditions of inquiry.

Publishing every exploratory utterance can discourage revision, expose personal information, and make a tentative statement appear to be an endorsed position. Selective documentation can identify the earlier formulation, the challenge, the resulting change, and the remaining obligation. Selection itself should be visible when it affects an account of the process.

Correction should address the relevant layer. An erroneous source attribution requires a source correction and review of dependent claims. An invalid proof requires a repaired argument or withdrawal

of the proposition. An inaccurate discovery narrative requires a historical correction even if the final proposition remains sound. A later insight into the argument can improve present understanding while leaving earlier competence uncertain.

The proposed responsibility is therefore differentiated and revisable. Its content follows the claim, the invited reliance, the actor's control, and the available response. It supports serious public discussion while leaving room for privacy, supported research, finite projects, and legitimate disagreement about the norms themselves.

12 Open Questions and Evaluation Designs

This section identifies research needed to assess and refine the proposed method. It separates conceptual questions about integration from empirical questions about revision practices and institutional questions about burden. The suggested designs are proposals for future investigation, with their evidential limits stated in advance.

12.1 Comparative assessment of revision episodes

A revision episode provides a possible unit of study when its initial claim, available materials, consequential challenges, and resulting formulation can be documented. The choice of unit should preserve the actual intellectual task and distinguish it from message or word counts.

One exploratory comparison could give participants matched philosophical problems and a common set of verified source excerpts. Conditions might vary the form of assistance: unrestricted dialogue, a process organized around claim and dependency records, or ordinary collaborative discussion. Any assignment procedure would need to address prior expertise, task difficulty, time, access needs, and interaction between participants. A small pilot could first test whether the tasks and evaluation criteria are intelligible.

Possible outcomes include accurate source attribution, identification of a consequential missing premise, appropriate narrowing of a claim, and retention of unresolved objections. Evaluators would need reasons for treating a revision as appropriate. Independent assessment can improve scrutiny, while disagreement among assessors may reveal an ambiguity in the task or criterion. A numerical agreement statistic would describe the assessment process, leaving substantive correctness open to further argument.

The number of revisions is a particularly weak stand-alone measure. A sound initial argument may need little change. Repeated stylistic changes may leave the central error untouched. An episode can also improve through justified resistance to a proposed revision. The target is the relation between reasons and changes, including reasoned retention.

12.2 Integration over time and across support conditions

The working account of epistemic integration requires further conceptual and empirical development. Its evaluation should distinguish current supported performance, subsequent retention, application to another problem, and accurate memory of discovery.

Open Problem 12.1 (Role-sensitive integration). Which capacities provide sufficient evidence of epistemic integration for different research roles, and how should appropriate support conditions alter that judgment without concealing consequential dependence?

A study could revisit the same participant after an interval and ask for an explanation of a vulnerable premise or an application to a nearby case. The later task would need a declared support regime. Removing every tool changes the target to unaided performance; allowing the original record measures a different supported capacity. Both may be informative when their purposes are explicit.

The intervention can itself teach. Discussion during an assessment may generate the understanding that the assessment seeks to measure. Researchers would need to distinguish evidence collected before

and after that exchange and avoid projecting the later result backward. Longitudinal evaluation also raises questions about attrition, changed interests, and the continuing value of the original research problem.

12.3 Critical diversity and the costs of the method

The protocol's value depends partly on the kinds of criticism it makes available and the labor it requires. These questions concern actual practice and cannot be settled by describing an ideal sequence.

Open Problem 12.2 (Consequential critical diversity). Under which conditions do additional interlocutors reveal materially different assumptions, source interpretations, or counterexamples, and when do their outputs reproduce a shared framing?

A comparison could trace the source and target of each objection and examine what diversity those relations reveal. An objection that changes the interpretation of the research question may contribute differently from several objections to one peripheral example. The proposed categories would themselves require refinement through cases.

Recording revisions, reading sources, and inviting criticism consume time. These costs can be justified when they prevent consequential error, but a rigid requirement can burden exploratory or low-stakes work. Evaluation should therefore consider the importance of the claim, the likely uses, available expertise, privacy, and the opportunity costs of documentation. A useful method should support proportionate choices and legitimate stopping.

12.4 Antecedent coverage and conceptual revision

The literature comparison remains selective. Further work may identify closer antecedents, alternative theories of understanding, or established practices that already implement parts of the proposal.

Such findings can change the paper's terminology and contribution. The current aim is to articulate a usable arrangement of retained insights and explicit safeguards for discussion. Conceptual priority remains unclaimed. The strongest future test is whether the arrangement improves the intelligibility and corrigibility of actual inquiry while preserving its openness to questions that the method itself has yet to anticipate.

13 Conclusion

This section brings together the paper's methodological contribution, its worked result, and the obligations that remain open. The proposed approach organizes AI-assisted philosophical inquiry around revisable claims, identifiable reasons, and appropriate contact with sources, objects, and other participants.

The method separates three evaluative questions. Discovery provenance concerns the history that available records support. Argumentative warrant concerns the reasons for a claim. Epistemic integration concerns a participant's ability, under declared supports, to explain and responsibly use those reasons. Each relation can matter, and evidence for one leaves the others in need of their own assessment.

The worked reconstruction shows the value of that separation. A retained distinction between formation and maintenance is used to construct a candidate temporal inference. The formal test establishes that a historical formation premise leaves a current exchange requirement logically open. Adding a bridge can restore validity while merely restating the disputed requirement, so an explanatory account needs independent content. A verified source passage also corrects an overbroad portrayal of a recognition-theoretical antecedent. The logical revision and the attributional revision contribute through different routes.

The proposed protocol gives a place to exploration, source verification before use, serious alternatives, consequential challenge, reasoned revision, and stabilization for further criticism. Its functions can recur in different orders. Formalization contributes where a local construction clarifies an inference; philosophical and empirical obligations remain attached to the interpretation of that construction.

AI assistance can participate in several functions of this process. Its outputs gain epistemic significance through the reasons, evidence, and responsibilities that connect them to an inquiry. Generated agreement, adversarial tone, and fluent prose require proportionate scrutiny, just as human contributions do. Accurate attribution and appropriate disclosure support that scrutiny without requiring complete public exposure of exploratory work.

The paper offers a conditional method for public and scholarly discussion. Evaluation of actual revision episodes, role-sensitive integration, critical diversity, and documentation burdens remains open. Those inquiries can refine the method and reveal circumstances in which another arrangement better serves the philosophical work.

DRAFT

References

- [1] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics, 2020.
- [2] John Dewey. *How We Think*. D. C. Heath, 1910. Selected chapters 1 and 6; Project Gutenberg ebook 37423: <https://www.gutenberg.org/ebooks/37423>.
- [3] Axel Honneth. *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Polity Press, 1995. Translated by Joel Anderson; pp. 104–105 consulted.
- [4] Helen E. Longino. *The Fate of Knowledge*. Princeton University Press, 2002. Introduction, pp. 8–9, consulted in publisher preview.
- [5] Murray Shanahan. Still “talking about large language models”: Some clarifications. arXiv:2412.10291v1, 2024.
- [6] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. arXiv:2310.13548v4, 2025. Inspected revision dated 10 May 2025 of work published at ICLR 2024 and first posted in 2023.